T-Shirt Size Estimator using Monocular Camera

Mukesh Cheemakurthi
Northeastern University
cheemakurthi.m@northeastern.edu

Ksheeraj Prakash
Northeastern University
prakash.ks@northeastern.edu

Abstract

One of the challenges customers face when shopping on e-commerce platforms is finding the right T-shirt size. Standard size charts generally include sizes like S, M, L, and XL. Even though these sizes appear standardized; the actual measurements can vary across brands. Given these challenges, it can be difficult for buyers to manually measure their size with a tape measure. This project investigates a system for estimating body measurements, such as shoulder-breadth and torso height, from single photographs using three approaches: a baseline method, a geometric method, and a deep learning (DL) method. The baseline method relies on MediaPipe pose detection to compute measurements based on key landmarks of the body, offering computational simplicity but struggling with inaccuracies caused by perspective distortions and pose variability. While, the geometric method enhances accuracy by incorporating person segmentation and clothing segmentation to identify landmarks with greater precision. Additionally, it employs a camera capture application to guide users in capturing images with minimal perspective error, ensuring robustness under ideal conditions. This method uses perspective correction and a reference object for accurate scaling, though it depends on the presence of a reference object in the image. The DL method uses a neural network to model complex relationships between pose landmarks and measurements, leveraging data augmentation to handle pose and lighting variability. This paper explores the strengths and limitations of each method, comparing their performance to determine the most effective approach for achieving precise body measurements. These findings aim to develop an accurate and user-friendly solutions for applications like e-commerce T-shirt size recommendations.

Introduction

I. Problem Definition

To select the correct T-shirt size is very challenging in the e-commerce industry, mainly due to inconsistencies in size standards across brands and the varying body measurements of individuals. These standard sizing charts often fail to account for these small variations which leads to a poor fit along with customer dissatisfaction, and an increase in product returns. Moreover, manual methods of taking body measurements, such as using a tape measure are inconvenient and are inaccurate especially when users attempt to measure themselves. These challenges highlight the need for an automated and user-friendly solution to recommend accurate T-shirt sizes based on body measurements.

II. Motivation

This project aims to bridge this gap by exploring three methods for predicting body measurements using images: Landmark Method(Baseline) , Geometric Method, and Deep Learning Method. The Baseline method uses pose landmarks to estimate body dimensions based on the relative positions of key points such as shoulders, chest, and hips. Whereas the geometric method builds on the baseline by incorporating corrections for perspective and scale using a reference object with known dimensions(height). This ensures a much reliable measurements even in non-standard conditions, such as skewed camera angles. Finally, the deep learning-based method uses pose landmarks and user-provided height as inputs to a neural network. This method learns complex, non-linear relationships between input features and target measurements, offering greater accuracy and robustness compared to the other approaches. By evaluating and comparing these three methods, this project seeks to identify the most accurate and scalable approach for predicting key body dimensions.

III. Open-Source Code Disclosure

This project was developed using open-source libraries and datasets in Table 1, these were integral to the implementation of all three methods:

Table 1: Open-source libraries and Datasets

Computer Vision:	• MediaPipe Pose: It's used across all models for human pose estimation and keypoint detection. • YOLOv8x-seg: It's applied in the geometric model for object detection and segmentation. • OpenCV: It's used in all models for basic to advanced image processing tasks.
Deep Learning:	• PyTorch: It employed exclusively in the DL model for deep learning tasks, including model training and inference.

	• scikit-learn: It's utilized in DL model for ML algorithm integration and evaluation.
Data Processing:	• NumPy: Its crucial for numerical calculations and data transformations in all models. • Pandas: It's used in geometric and DL models for efficient data wrangling.
Visualization:	• Matplotlib: It's used across all models for creating visualizations. • Seaborn: It's utilized in the DL model for enhanced statistical plotting and data analysis.
BodyM Dataset	A dataset containing paired silhouette images and corresponding anthropometric measurements, including height and shoulder breadth. It's used to train, test, and validate the models.
NOMO-3D-400-Scans Dataset	A dataset comprising 3D body measurements. It's used to train, test, and validate as well as for fine-tuning the models.

Related Work

Our project combines advanced techniques in object detection, deep learning, and pose analysis to create an automated system for accurate T-shirt size prediction from single images. Modern advancements in computer vision have enabled significant progress in automated body measurement systems, crucial for applications such as personalized clothing recommendations. Zhou et al. [1] introduced YOLO-MDE, which integrates object detection with monocular depth estimation to enhance the precision of body measurements by adding a 3D perspective to 2D image analysis. Similarly, Cai et al. [2] demonstrated that depth-enhanced object detection significantly improves spatial accuracy, highlighting its relevance for body dimension estimation. Pose estimation is central to our approach, as it enables accurate identification of key body landmarks. Sun et al. [3] developed high-resolution networks to detect critical landmarks such as shoulders, torsos, and hips, providing the basis for precise measurement predictions. While, Graving et al. [4] introduced DeepPoseKit, a lightweight deep-learning framework for efficient and accurate keypoint detection, which is integral to extracting pose-based features. Also, Omran et al. [5] further enhanced body measurement accuracy through Neural Body Fitting (NBF), a model combining segmentation with statistical body shape modeling, enabling reliable predictions of body dimensions. These real-world applications demand robustness against distortions caused by factors such as poor lighting or camera angles. Whereas, Zhang et al. [6] proposed data augmentation and perspective correction techniques to mitigate these challenges, aligning with our adoption of perspective correction for accurate measurements. Additionally, Gadhiya and Kalani [7] validated the efficiency and precision of MediaPipe for pose detection, reinforcing our choice of this tool for identifying body landmarks. To convert pose-based pixel measurements into real-world dimensions, depth estimation is indispensable. Poudel et al. [8] reviewed monocular depth estimation methods, emphasizing their ability to provide 3D information from single images without expensive hardware, a feature that informs use in our project. By leveraging these methodologies, our project integrates YOLO for object detection, MediaPipe for landmark detection, and perspective correction into a streamlined system that addresses challenges like distortion correction, measurement accuracy, and real-world variability. This unified framework enables precise and scalable body measurement predictions, providing a transformative solution for accurate online T-shirt size recommendations, improving user experience, and advancing automated clothing systems.

Methods

This project is implemented using three distinct methods for predicting body measurements: the Landmark-Based Method (Baseline), the Geometric Method, and the Deep Learning (DL) method. Each method is designed to extract key body dimensions, including shoulder breadth, and shoulder-to-crotch length, from images. The baseline method uses simple pose landmarks, the geometric method incorporates perspective correction and scaling, and the DL method uses advanced neural networks to model complex relationships. Each method is described in detail below.

I. Baseline Landmark-Based Method

The baseline method focuses on detecting key body landmarks using MediaPipe, a tool for pose estimation. The primary input is an image of the user and their height in centimeters. MediaPipe identifies anatomical landmarks, such as shoulders, hips, and the nose, as x and y coordinates, allowing for the calculation of key body measurements: shoulder width and shoulder-to-crotch length(Figure 1). The method calculates shoulder width as the horizontal distance between the landmarks for the left and right

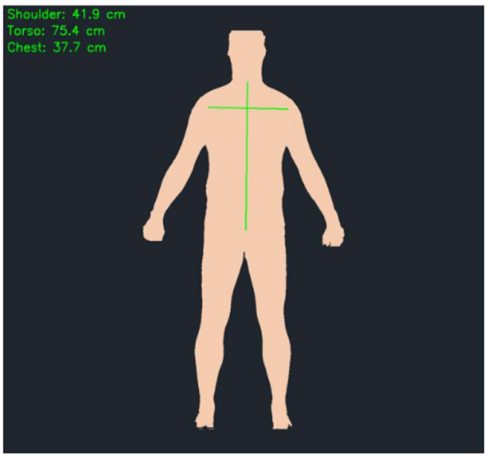

Figure 1: Visualization of Baseline Model

The shoulder-to-crotch length is determined by identifying the midpoint between the shoulder and nose landmarks (representing the neck) and calculating the vertical distance from this point to the midpoint of the hip landmarks. These pixel measurements are then scaled to real-world dimensions using the user's height and a derived scaling factor. A scaling factor is computed based on the pixel height of the user (distance from the nose to the ankle) and their real-world height in centimeters. This factor translates pixel measurements to real-world units. The flowchart of the algorithm is given in Figure 2.

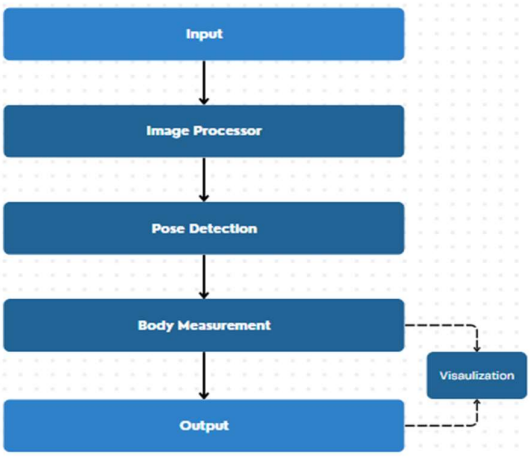

Figure 2: Flow Diagram of the Baseline-Landmark Method

Algorithm

Input:

- Input the image path and user height in centimeters.

Image Processor:

- Load the input image.
- Preprocess the image (apply Gaussian blur, thresholding, or RGB conversion as needed).

Pose Detection:

- Use MediaPipe to extract pose landmarks.
- Validate the presence of required landmarks.

Body Measurement:

- Calculate ***Shoulder Width*** as the horizontal distance between shoulder landmarks.
- Calculate ***Shoulder-to-Crotch Length*** as the vertical distance between the midpoint of neck and the hip. Fig 2: Flow Diagram of the Geometric Method
- Scale these pixel measurements to real-world dimensions using the user height-derived scaling factor.

Output:

- Return the measurements.
- Optionally visualize the landmarks and measurements on the input image.

II. Geometric Method

The geometric method enhances the baseline method by incorporating perspective correction and scaling factors derived from a reference object included in the image. This method addresses challenges such as camera angle distortions and ensures accurate conversion of pixel dimensions into real-world measurements. A reference object, such as a phone or card with known dimensions, is detected and used to calculate the pixel-to-real-world scaling factor or if unavailable the person height can be used as a reference point(Figure 4). This factor is applied to measurements such as shoulder breadth and shoulder-to-crotch length. The image undergoes perspective correction, which aligns the plane of the image with the body plane, reducing distortions caused by non-ideal angles. The combined use of scaling and correction ensures accurate measurements regardless of the user's posture or the camera angle. Finally, the T-shirt segmentation from the image is achieved by training a YOLO model to detect the T-shirt. This process is similar to torso segmentation but focuses specifically on detecting the T-shirt and finding the torso height. The flowchart of the algorithm is given in Figure 3.

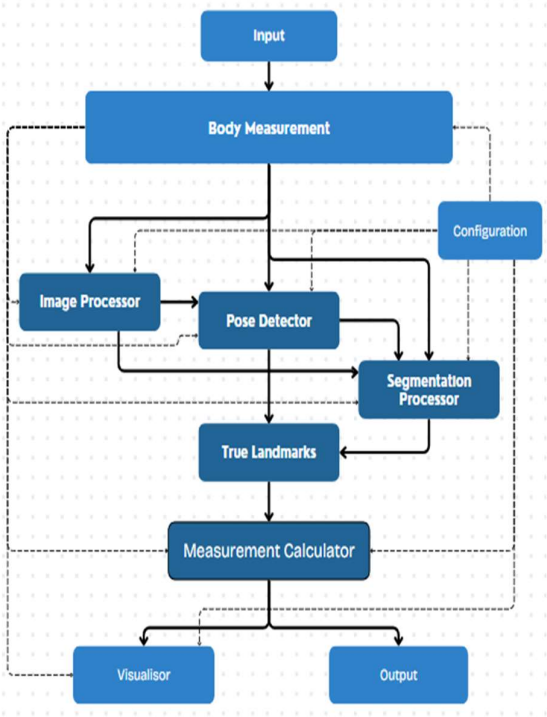

Figure 3: Flow Diagram of the Geometric Method

Algorithm

Input:

- Accept the image path and user-provided height in centimeters.

Configuration:

- Initialize settings for models and processors

(e.g., YOLO and MediaPipe).

Image Processor:

- Preprocess the input image using CLAHE and denoising.
- Perform segmentation with the reference object.

Pose Detection:

- Use MediaPipe to detect anatomical landmarks like shoulders, hips, and nose.
- Validate and refine extracted landmarks.

Segmentation:

- Detect the reference object and calculate its pixel dimensions.

True Landmarks:

- Adjust and confirm the extracted pose landmarks relative to the segmented object.

Measurement Calculator:

- Perform perspective correction to align the image plane.
- Calculate pixel distances for features like shoulder breadth and shoulder-to-crotch length.
- Apply the scaling factor derived from the reference object to convert measurements to real-world units.

Output:

- Return the calculated measurements in a structured dictionary.
- Generate visualizations showing the overlay of pose landmarks and measurements.

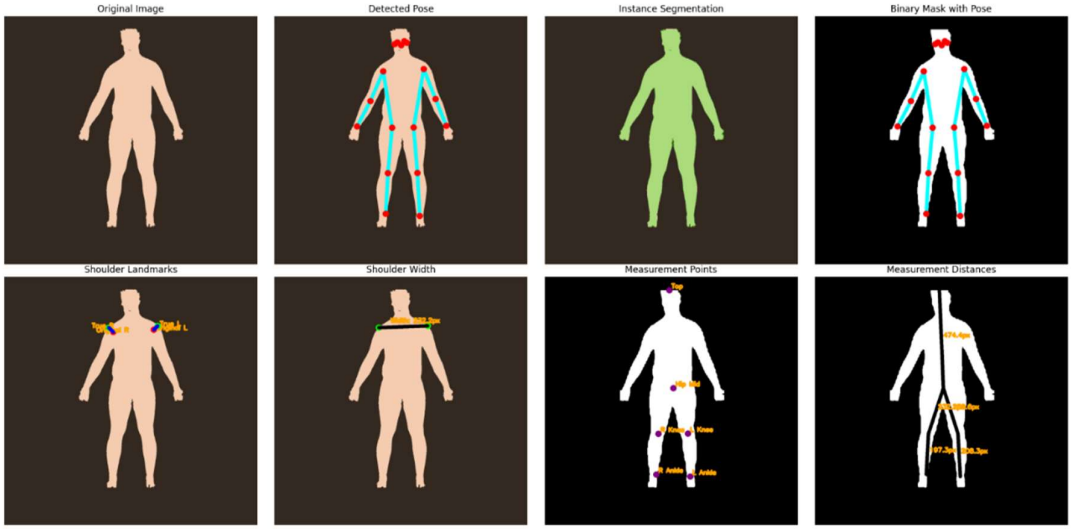

Figure 4: Visualization of the Geometric Model

The geometric method significantly improves the accuracy of body measurement predictions by addressing key limitations of the baseline approach. It builds a robust pipeline capable of processing complex input scenarios, paving the way for better user experience and higher reliability.

III. Deep Learning Method:

The Deep Learning (DL) method represents the most sophisticated approach, using neural networks to extract meaningful patterns from pose landmarks and user-provided height. This approach captures relationships between pose landmarks and body measurements, making it useful to variations in body posture, lighting, and camera angle. The input features include Pose landmarks extracted by MediaPipe and User-provided height for real-world scaling. The architecture comprises a shared encoder and task-specific prediction heads for shoulder breadth, chest circumference, and shoulder-to-crotch length. A shared encoder processes the input features through layers of fully connected neurons with Rectified Linear Unit (ReLU) activation and batch normalization to enhance feature extraction and generalization. Each prediction fine-tunes the specific measurement computation using specialized weights. During training, data augmentation techniques such as random scaling, translations, and Gaussian noise are applied to simulate real-world distortions and improve the model's robustness. The model uses Mean Squared Error (MSE) as the loss function and optimizes parameters using the Adam optimizer with early stopping to avoid overfitting. The flowchart of the algorithm is given in Figure 5.

Algorithm

Input:

- Receive image path and user-provided height.

Image Processing:

- Enhance the image using CLAHE and denoising techniques.

Landmark Detection:

- Use MediaPipe to detect key body landmarks (e.g., shoulders, hips, knees).
- Standardize the landmarks and configure them for consistency.

Dataset Preparation:

- Create the training dataset using standardized landmarks.
- Apply data augmentation techniques (random scaling, translations, and Gaussian noise).

Model Training:

- Use a shared encoder with task-specific heads for predicting shoulder breadth, chest circumference, and shoulder-to-crotch length.
- Train the model with MSE loss and optimize using the Adam optimizer.
- Perform *Backward Pass* to update model weights iteratively.
- Apply early stopping to prevent overfitting.

Prediction:

- Conduct a *Forward Pass* for inference on new images.

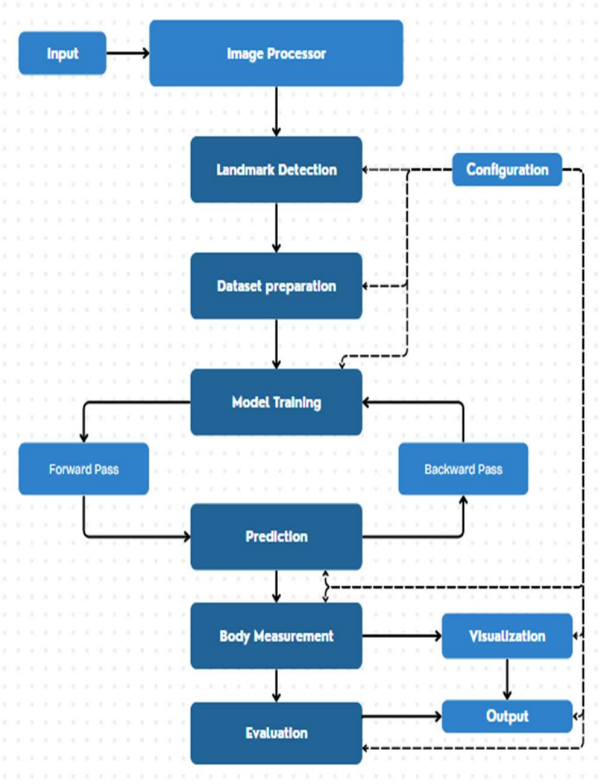

Figure 5: Flow Diagram of the Deep Learning Method

- Predict real-world measurements using the trained model and scaling factors derived from user-provided height.

Body Measurement and Visualization:

- Compute body measurements from model outputs.
- Generate visual overlays on the input image to highlight pose landmarks and measurements.

Evaluation and Output:

- Evaluate performance metrics (e.g., MAE, RMSE) for predicted measurements.
- Output measurements as structured dictionaries and visualizations.

This method relies on neural networks allowing the system to generalize well by achieving high accuracy, even with variations in input data. However, this approach requires significant computational resources for training and deployment.

Experiments

These experiments were conducted on a dataset specifically curated for body measurement tasks. The dataset consisted of images with corresponding ground truth measurements for three key dimensions: shoulder breadth, chest circumference, and shoulder-to-crotch length. Each image was associated with real-world measurements collected using calibrated tools. The dataset was split into training (80%) and testing (20%) sets, with the testing set further divided into standard perspective poses (e.g., standing upright) and challenging perspective poses (e.g., distorted angles).

The project is designed to answer the following questions:

1. ***Accuracy*:** Which method provides the most accurate predictions for body measurements under standard conditions?
2. ***Robustness*:** How does each method handle challenging conditions, such as extreme angles, lighting distortions, or occlusions?
3. ***Performance Comparison*:** Can the Deep Learning (DL) approach significantly outperform the baseline landmark-based and geometric methods in terms of both accuracy and robustness?
4. ***Scalability*:** How feasible is each method for real-world applications in terms of computational requirements and dependence on additional inputs, such as reference objects?

Experimental Setup

The experimental setup checks the Baseline Landmark-Based Method, Geometric Method, and Deep Learning Method approaches to compare their accuracy, robustness, and computational efficiency in predicting shoulder breadth and shoulder-to-crotch length. The evaluation process is as follows

1. **Metric Evaluation:**
   Mean Absolute Error (MAE): Measures the average absolute difference between predictions and ground truth values.
   Root Mean Squared Error (RMSE): Emphasizes larger errors by computing the square root of the average squared error.
   Accuracy (%): Proportion of predictions within a ±2cm threshold of the ground truth measurements.
2. **Visual Results:** Pose detection and landmark extractions are used to verify the consistency of pose estimations across methods. Scatter plots demonstrate the accuracy of predictions compared to ground truth, highlighting deviations for non-perspective and perspective-corrected setups. Training curves for the Deep Learning method illustrate in Figure 6, the reduction in error over epochs and the model's convergence.

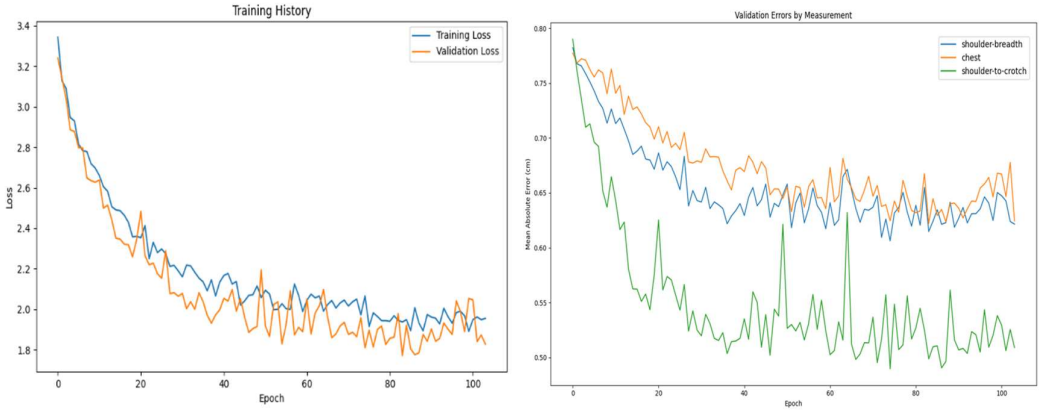

Figure 6: Training Graph(L) and Validation Error Graph(R) of DL Model

3. **Comparison Framework:** The Baseline Landmark-Based Method serves as the foundational benchmark, offering the challenges posed by uncorrected measurements and the lack of geometric adjustments. The Geometric Method builds upon this by incorporating instance segmentation significantly enhancing robustness against variations in land mark detection and user posture. Finally, the Deep Learning Method leverages advanced neural networks to model complex relationships between pose landmarks and body dimensions, achieving superior accuracy and adaptability.

Results and Observation

Baseline Landmark-Based Method

This method relies solely on pose landmarks detected using MediaPipe, without any geometric corrections or advanced scaling mechanisms. Measurements were scaled using the user-provided height.

Performance: This method exhibited the highest error rates and lowest accuracy across all metrics. For instance, the RMSE for shoulder breadth predictions was 7.094 cm, and the accuracy was only 33.97% as shown in Figure 7. The method struggled with challenging poses, as the image were distorted or captured with skewed angles.

Geometric Method

This approach extends the baseline by introducing perspective correction and scaling based on a known reference distance. Geometric transformations align the body plane to correct distortions caused by camera angles, improving measurement accuracy.

Performance: The geometric method showed substantial improvements over the baseline. By incorporating perspective correction and reference-based scaling, it reduced the MAE for shoulder breadth from 5.549 cm

(baseline) to 2.749 cm (non-perspective) and further to 2.433 cm (with perspective correction) as shown in Figure 8.

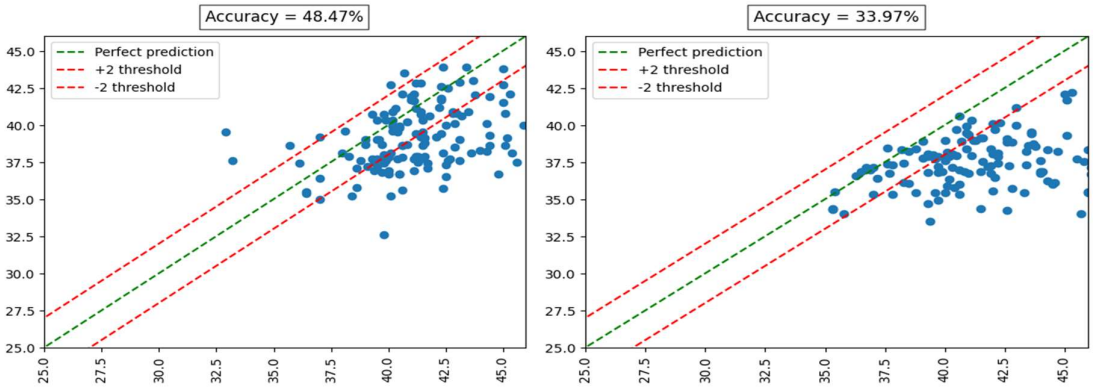

Figure 7: Accuracies of Baseline Method – Perspective Images(L) and Non-Perspective Images(R)

Robustness: The perspective correction improved accuracy under distorted poses, achieving an overall accuracy of 80% for shoulder breadth predictions. However, in cases where the reference object was not correctly identified or was missing, the method's performance declined.

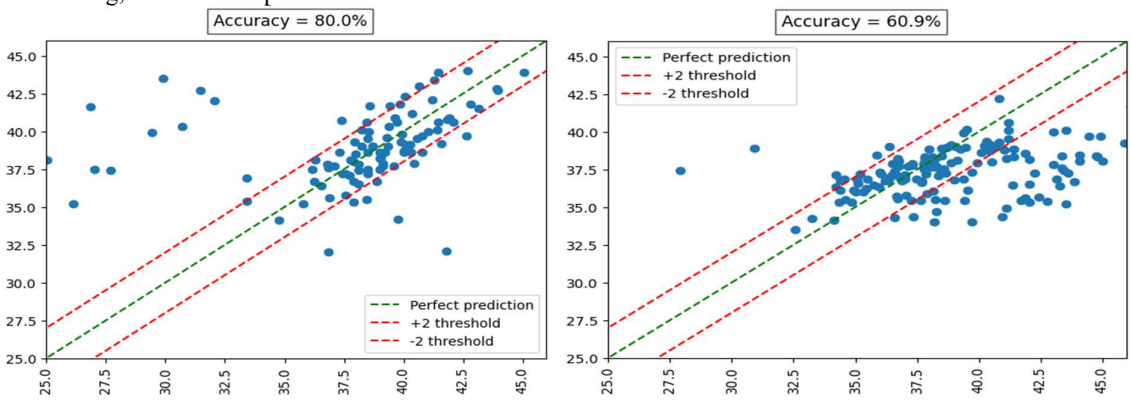

Figure 8: Accuracies of Geometric Method – Perspective Images(L) and Non-Perspective Images(R)

Deep Learning Method

This approach leverages a neural network trained on pose landmarks and user height to model complex, non-linear relationships between landmarks and body dimensions. Data augmentation techniques, such as random scaling, rotation, and noise injection, were employed to improve generalization under diverse conditions

Performance: The DL method consistently achieved the better results. For shoulder breadth, it achieved MAE of 1.166cm and an RMSE of 1.498cm, with an accuracy of 89.17% (non-perspective) as shown in Figure 9. Its ability to model non-linear relationships between landmarks and body dimensions is a significant advantage.

Robustness: The DL method performed well even under challenging conditions, owing to extensive data augmentation during training. However, introducing perspective correction slightly reduced its accuracy to 75.98%, likely due to an over-reliance on augmented data that assumed uniform conditions.

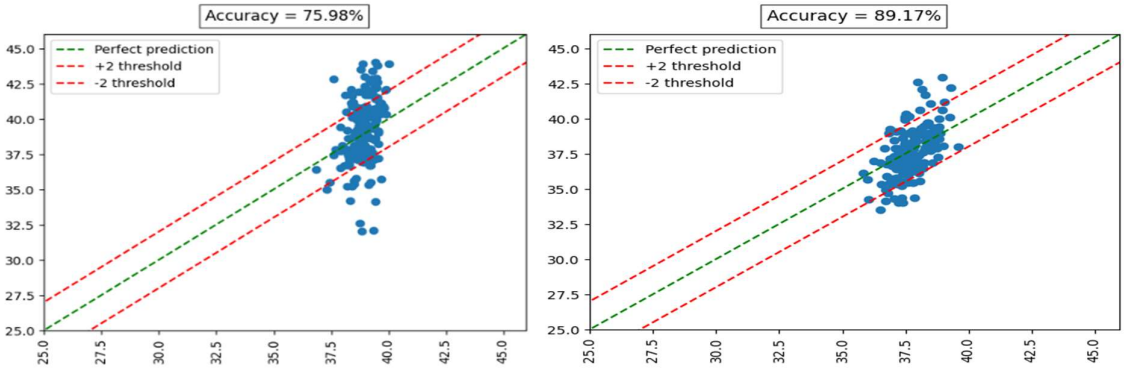

Figure 9: Accuracies of DeepLearning Method – Perspective Images(L) and Non-Perspective Images(R)

The results are summarized in the table below:

Table 2: Results of all models

Model Type	Approach	RMSE	MAE	Accuracy (%)
Baseline Models	Non-Perspective	7.094	5.549	33.97
	Perspective	9.089	4.985	48.47
Geometric	Non-Perspective	3.758	2.749	60.90
	Perspective	4.037	2.433	80.00
Deep Learning	Non-Perspective	1.498	1.166	89.17
	Perspective	2.182	1.685	75.98

Strengths and Limitations

Table 2: Strengths and Limitations of all three models

Model	Strengths	Limitations
Baseline Method	Computationally efficient and straightforward.	Lacks precision and robustness under perspective distortions and non-standard poses.
	Effective for detecting measurements in standard poses.	Dependent on accurate pose landmark detection without geometric corrections.
Geometric Method	Performs better under perspective distortions due to geometric correction.	Require reference object in the image, which may not always be practical
	Scales pixel dimensions to real-world measurements effectively.	More computationally intensive compared to the baseline method.
Deep Learning Method	Provides the highest accuracy for non-perspective images by leveraging complex neural networks.	Performance decreases in perspective images due to a training bias on non-perspective datasets
	Robust to variations in lighting, pose, and camera angles through data augmentation.	Requires significant computational resources for training and inference.
	Scalable for additional measurements with minimal reconfiguration	Implementation and debugging are more complex than simpler methods.

Conclusion

The experimental results reveal that while the baseline method is computationally efficient and suitable for standard poses, it lacks precision and robustness in handling perspective distortions. The geometric method, with its incorporation of perspective correction and geometric scaling, performs significantly better in scenarios with distorted or non-standard poses, making it the optimal choice for perspective images. On the other hand, the deep learning method achieves the highest accuracy and robustness for non-perspective images, as it effectively models complex relationships between pose landmarks and body dimensions. However, its performance declines in perspective images due to its training bias toward non-perspective scenarios and reliance on consistent data augmentation. Overall, the geometric method is best suited for perspective images, while the deep learning method is preferable for non-perspective conditions, highlighting the potential for a hybrid approach to leverage the strengths of both methods.

Future Directions

- Integration with Depth Estimation: Incorporating depth estimation models, such as Midas, can eliminate the need for reference objects while improving accuracy.
- Hybrid Models: Combining the geometric method with deep learning could enhance scalability and robustness under diverse conditions.
- Real-Time Implementation: Optimizing the DL model for deployment on edge devices will enable real-time predictions in resource-constrained environments.
- 3D Image Reconstruction: Reconstruction 3D Object from images captured from different angles using a monocular camera

References

[1] Zhou, T., et al. YOLO-MDE: Object detection meets monocular depth estimation. *International Journal of Computer Vision* (2022).

[2] Cai, L., et al. Unified object detection and depth estimation for autonomous systems. *IEEE Transactions on Robotics* (2022).

[3] Sun, K., et al. High-resolution networks for human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[4] Graving, J. M., et al. DeepPoseKit: A toolkit for fast and accurate pose estimation. *Nature Methods* (2019).

[5] Omran, M., et al. Neural Body Fitting: Combining CNNs and statistical models. *Proceedings of CVPR* (2018).

[6] Katircioglu, I., et al. Structured prediction for pose and shape estimation. *NeurIPS* (2018).

[7] Gadhiya, R., & Kalani, K. Comparative review of pose estimation architectures. *Journal of Computer Vision* (2021).

[8] Zhang, Y., et al. Overcoming distortions in image-based body measurements. *Pattern Recognition Letters* (2023).

[9] BodyM Dataset. (2020). Hosted on AWS Open Data Registry. Available at: https://registry.opendata.aws/bodym/

[10] Yan, S., Wirta, J., & Kämäräinen, J. Anthropometric clothing measurements from 3D body scans. *Machine Vision and Applications*, 31(7), 2020. DOI: https://doi.org/10.1007/s00138-019-01054-4